

一种实现搜索引擎个性化服务的方法

A Method of Providing Individual Service in Search Engine

谭琼^{1,2} 李晓黎² 史忠植²

(中国科学技术大学管理学院 北京 100080)¹(中国科学院计算技术研究所 北京 100080)²

Abstract With the rapid increase of the information on the Internet, it is more and more difficult for people to select the needed information. To solve this problem, this paper presents a method to provide individual service in search engine. This method adopts the theory of statistics and can get a describing file of user's interest by a learning algorithm. It only takes a little man-machine interaction. The file is the base of providing individual service. Furthermore, in order to track the variation of user's interest dynamically, this paper uses the local autonomous Agents to percept the user's action.

Keywords Information retrieval, Agent, Keywords, Interest degree

1 引言

随着 Internet 网上信息的剧增,人们在利用计算机检索信息时变得越来越困难。一方面是信息非常丰富,收到太多的信息却无从选择和消化;另一方面是信息迷失,人们难于找到自己真正所需的信息。原因是:当前的信息检索技术很少考虑用户的个人兴趣。由于不同的用户有不同的需求,所以显然不能满足各种用户的需要。解决该问题的一个方法是:在准确、动态地描述用户的兴趣和偏好的基础上,使搜索结果呈现出个性化的特点。因此,如何描述和跟踪用户的兴趣和偏好就成为网上信息检索的关键。它既可以协助计算机准确定位用户之所需,而且也成为实现主动信息服务的先决条件。

目前,用户兴趣的描述有的直接采用人为描述其兴趣文件来实现。这种做法的描述粒度不好把握,同时由于用户不熟悉领域知识,因此用户不易精确地表达其兴趣。此外,由于兴趣的广泛性与易变性,使用静态的方式表达来描述兴趣是不能在信息检索中发挥大的作用的。

本文提出了一种用户不直接编辑兴趣描述文件的动态兴趣学习方法。该方法采用统计学的原理,仅用较少的人机交互,即询问某个用户对一定数量网页的标题是否感兴趣,就可通过兴趣学习算法自动获得该用户初始的个性化兴趣描述文件,从而用于用户的相关信息检索,并基本保证符合用户的兴趣。这种方法的一个好处是:可以人为地控制交互的内容。

为了进一步精确描述用户的兴趣或动态跟踪用户兴趣的改变,我们采用本地自主感知主体(Agent)来对用户的行为进行感知,实时对用户兴趣进行监测并获取用户感兴趣的相关信息及其感兴趣的程度。譬如用户访问时的停留时间、访问次数、保存、编辑、修改等动作。同时用户输入的关键词也作为动态更新兴趣文件要考虑的因素。此外,还要考虑噪声的剔除问题,避免偶然的情形。

本文介绍了特征词的分类及相关概念;提出了用户兴趣学习和识别算法;设计了用户兴趣源的动态获取和更新;最后给出结论。

2 特征词的分类及相关概念

对于一篇文章或一个网页而言,人们一般在对标题阅读后就可以判断是否对其感兴趣。这是因为标题是概括文章中心的句子,它言简意赅地表达了文章的核心内容。有人抽样统计,网页标题与文章主题的基本符合率高达 80%。通过对标题的学习,我们可以把握文章的中心内容,减少了用户看全文的负担。同时只利用较少的信息学习,加快了学习的速度。

标题的重要性主要由网页标题中的构成成员即特征词的性质决定。经过仔细分析可知,特征词可以分成三类。一类是积极特征词:由于它的存在使用户对网页感兴趣;第二类为消极特征词:由于它的存在用户对此不感兴趣;第三类是无用特征词(严格说不是特征词):它的存在不作为判断用户有无兴趣的依据。

下面给出有关特征词、标题的特征程度的一些定义。

定义 1 设网页标题集为 $\{title_i\}$, 其中 $i=1,2,\dots,k$ (k 为标题集的元素个数), 停用词库为 $stoplist$, w_{ij} 为 $title_i$ 的第 j 个特征词, 其中 $j=1,2,\dots,n$ (n 为特征词的个数), $p\{title_i\}$ 表示用户对 $\{title_i\}$ 感兴趣的概率, $p\{title_i, w_{ij}\}$ 表示当 $title_i$ 的第 j 个词为 w_{ij} 时用户感兴趣的概率, 则当给定的一个较小阈值 $\varepsilon > 0$, 且 $w_{ij} \notin stoplist$ 时:

若 $|p\{title_i, w_{ij}\} - p\{title_i\}| < \varepsilon$, 则 w_{ij} 为无用特征词;

否则 若 $p\{title_i, w_{ij}\} > p\{title_i\}$, 则 w_{ij} 为积极特征词;

若 $p\{title_i, w_{ij}\} < p\{title_i\}$, 则 w_{ij} 为消极特征词。

定义 2 特征程度:

$$\begin{aligned} Degree(p\{title_i, w_{ij}\}, p\{title_i\}) \\ = p\{title_i, w_{ij}\} \lg \frac{p\{title_i, w_{ij}\}}{p\{title_i\}} - \\ (1 - p\{title_i, w_{ij}\}) \lg \frac{(1 - p\{title_i, w_{ij}\})}{(1 - p\{title_i\})} \end{aligned}$$

Degree 既符合特征词的分类要求,同时其值反映了特征词对用户兴趣的影响程度。当词 w_{ij} 为积极、消极、无用特征词时, Degree 的值分别为正、负或近似零。

定义 3 令 t_i 为用户对 $\{title_i\}$ 集感兴趣标题的总数, t_u 为用户对 $\{title_i\}$ 集无兴趣标题的总数, 则 $p\{title_i\} = t_i / (t_i + t_u)$;

谭琼 在职研究生,研究方向为信息管理。李晓黎 博士研究生,主要研究方向为数据采掘、信息检索。史忠植 研究员,博士生导师,主要研究方向为人工智能、知识工程。

令 t_n 为词 w_i 出现在 $\{title_i\}$ 集时积极特征词总数, t_n 为词 w_i 出现在 $\{title_i\}$ 集时消极特征词总数, q 为稀疏度因子, 是衡量词的支持程度的一个阈值。

若 $t_n + t_n < q$, 则 $p\{title_i, w_i\} = 0$, 否则 $p\{title_i, w_i\} = t_n / (t_n + t_n)$

当 $t_n + t_n < q$, 说明词 w_i 出现次数太少, 不能作为判断兴趣的依据(新词也作同样处理)。

定义4 标题的兴趣度:

$$Interest(title_i) = \sum_{j=1}^k Degree(p(title_i, w_{ij}), p(title_i))$$

$Interest$ 反映了某个 $title_i$ 中全部特征词对用户兴趣的总体贡献程度。 $Interest > 0$, 表明用户对 $title_i$ 有兴趣。同时其值越大, 表明用户的兴趣程度越高。

3 用户兴趣学习和识别算法

兴趣的挖掘与应用可分为两个基本步骤。在训练阶段, 给出用户一定数量的网页标题后, 请用户判断是否有兴趣进行统计学习。学习算法由此给标题中的特征词归类并计算其特征程度, 构成兴趣描述文件, 作为兴趣刻画的基础。在应用阶段, 对任意给出的网页, 我们可以根据兴趣描述文件计算该标题的兴趣度, 从而可以判断用户对该网页是否有兴趣。

3.1 兴趣学习算法

兴趣描述文件中除了存放了用户名、用户的标识及其用户的其它信息外, 还存放了特征词 $word_i$, 它出现在感兴趣标题中的次数为 t_n , 出现在无兴趣标题中的次数为 t_n 。令 s 为兴趣描述文件中特征词总数, 对于训练集 $X = \{title_i\}$, 其中 $i = 1, 2, \dots, k$, 学习算法描述如下:

Step1 $m = 1, t_n = 0, t_n = 0, t_n = 0, t_n = 0, s = 0$, 打开训练集 X 。

Step2 若 $X = \emptyset$, 则算法结束。否则取一个 $title_m \in X$, 设其特征词序列为 $w_{m1}w_{m2}\dots w_{mn}$, $X = X - \{title_m\}$ 。

Step3 将 $title_m$ 显示给用户, 请用户判断是否感兴趣, 若不感兴趣, 则 go to Step7。

Step4 $t_n = t_n + 1$ 。

Step5 循环 $j = 1$ TO n 执行

对任意 $w_{mj} \in stoplist$ 在文件中查特征词字段是否已经存在, 若找到, 即存在 $word_i = w_{mj}$, 则 $t_n = t_n + 1$;

否则 $s = s + 1$, 将 w_{mj} 加入到文件中, 并令 $t_n = 1, t_n = 0$;

Step6 $m = m + 1$, go to Step2。

Step7 $t_n = t_n + 1$ 。

Step8 循环 $j = 1$ TO n 执行

对任意 $w_{mj} \in stoplist$ 在文件中查特征词字段是否已经存在, 若找到, 即存在 $word_i = w_{mj}$, 则 $t_n = t_n + 1$;

否则 $s = s + 1$, 将 w_{mj} 加入到文件中, 并令 $t_n = 0, t_n = 1$ 。

Step9 go to Step2。

3.2 兴趣判别算法

构成初始的用户兴趣文件后, 我们可以利用它来判断任意给定的网页集 $Y = \{page_i\}$, 该用户是否感兴趣。该算法描述如下:

Step1 打开网页集 Y , 得到其对应的标题集 $X = \{title_i\}$, 打开兴趣描述文件, 初始化稀疏度因子 q 。

Step2 若 $X = \emptyset$, 则 go to step5; 否则取一个 $title_i \in X$, 设其特征词序列为 $w_{i1}w_{i2}\dots w_{in}$, $X = X - \{title_i\}$ 。

Step3 读入兴趣描述文件中的 t_n 与 t_n , 计算 $p\{title_i\}$ 。

Step4 循环 $j = 1$ TO n 执行

对任意 $w_{ij} \in stoplist$ 在文件中查特征词字段是否已经存在, 若找到, 即存在 $word_i = w_{mj}$, 则若 $t_n + t_n < q$,

则 $p\{title_i, w_{ij}\} = 0$, $Degree(p\{title_i, w_{ij}\}, p\{title_i\}) = 0$

$Degree(p\{title_i, w_{ij}\}, p\{title_i\})$

$$= p\{title_i, w_{ij}\} \lg \frac{p\{title_i, w_{ij}\}}{p\{title_i\}} - (1 - p\{title_i, w_{ij}\}) \lg \frac{(1 - p\{title_i, w_{ij}\})}{(1 - p\{title_i\})}$$

$$Interest(title_i) = Interest(title_i) + \sum_{j=1}^k Degree(p\{title_i, w_{ij}\}, p\{title_i\})$$

否则 $p\{title_i, w_{ij}\} = t_n / (t_n + t_n)$

Step5 将 $Interest$ 的值按从大到小的顺序排序结果。

Step6 将 $Interest > 0$ 的网页提供给用户, 算法结束。

4 用户兴趣源的动态获取和更新

前面我们分别介绍了用户兴趣的辨识和学习方法, 它们都是以网页的标题为中心, 根据文章标题所涉及的关键字来获取用户的兴趣。但这些提交给用户回答的网页规模与分布可能很难恰当地控制, 而且用户对交互式方式的忍受程度也不好把握。这样不一定能反映用户的真正兴趣, 也不能跟踪用户兴趣的动态变化。另外, 由于用户兴趣的广泛性, 用手工收集用户兴趣源显然无法准确及时地反映用户的兴趣及其变化。因此, 我们采用了 Agent 技术来监测用户的行为, 通过实时获取用户感兴趣的相关信息及其感兴趣的感性程度来感知用户的兴趣。

我们将 Agent 设计成一个后台程序, 它在 Internet 浏览器的后台运行, 并时刻监视浏览器的当前状态以及用户在浏览器中的操作。Agent 的结构如下:

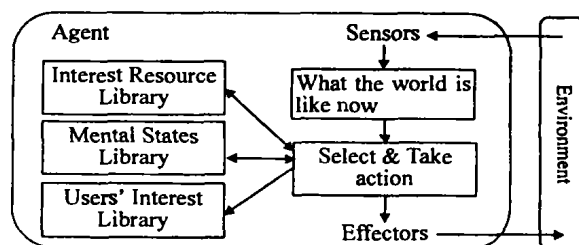


图1 Agent 结构示意图

Agent 的基本工作原理为: Agent 通过传感器(Sensor)感知得到环境的当前状态, 然后根据自身对外部世界的认识(如信念、约定等)以及自身的承诺, 选择与该状态相适应的行动规则, 即根据 Agent 自身的心智状态选择动作, 并采取相应的行动, 如从环境中获取用户的兴趣源, 利用用户兴趣动态分析算法更新用户兴趣等。另外, 主体在行动过程中, 会适时调整自身的心智状态(如信念、意图等), 以便适应新的环境要求。

其中, Agent 的传感器主要监控并感知用户在浏览器中的如下动作, 而 Agent 则根据这些动作作出相应的反应。

1) 页面的访问次数。访问同一页面的次数越多, 说明对该页面的内容越感兴趣。

当主体检测到浏览器的页面更新时, 首先获取页面的

URL 及标题,并搜索用户兴趣源数据库,计算得到该页面的访问次数。然后利用前面描述的兴趣判别算法计算出用户对该页面的最低感兴趣程度。最后将访问次数换算成特定的兴趣因子,得到用户对该页面的感兴趣程度。

2) 页面上的停留时间。当浏览器为活动窗口且系统处于活动状态时,对同一页面的关注越多即停留时间越长,表明对该页面的兴趣也越大。

在浏览器窗口为当前活动窗口的情况下,当主体发现浏览器的页面没有被刷新时,就会认定用户一直在关注当前页面,并开始计数,直到用户离开浏览器或访问其它页面为止。最后,主体将用户在同一页面上的驻留时间转换成对应的兴趣因子,并得到用户对该页面的感兴趣程度。

3) 页面内搜索字符串或关键字。一般来说,用户在页面内搜索的字符串或关键字都是用户感兴趣的东西。

主体在捕捉到搜索动作时,首先获取所搜索的字符串或关键字,并将该字符串看成是用户感兴趣的标题的一部分,然后调用兴趣判别算法计算并更新用户的兴趣。

4) 编辑、修改并保存页面。用户一般只会保存感兴趣的東西,而编辑或修改页面也往往是为了将所下载的页面更改成自己的页面,并会最终保存下来。

主体在捕捉到页面保存动作时,直接将用户对该页面的感兴趣程度加倍。

结束语 Internet 网上信息量过载已导致我们难以利用其巨大的价值。如何将繁多的信息按用户的兴趣从杂乱到有序是一个有意义的课题,这也是搜索引擎实现个性化服务的关键。本文提出一种学习算法,只用较少的人机交互就准确

地把握了用户的兴趣所在。Agent 技术使得我们根据用户的行为可以动态地对兴趣进行跟踪,在理论和实践上较适合于 Internet 上数量众多、经常改变的信息的情形。学习算法的动态性使系统容易跟踪用户兴趣的改变。

参考文献

- 1 Cheng Xueqi, Bai Shuo. A Decision-Tree-Based Mechanism for Information Difffluence[C]. In: Proc. 1998 Intl. Conf. on Chinese Information Processing. China, 1998
- 2 Aha W, Bankert L. A comparative Evaluation of Sequential Feature Selection Algorithms[C]. In: Proc. of AI and Statistics Workshop 1995
- 3 Domingos P. Context-Selection for Lazy learners[J]. Artificial Intelligence Review, 1997
- 4 Cost S, Salzberg. A Weighted Nearest Neighbor Algorithm for Learning with Sysbolic Features[J]. Machine Learning, 1997, 10: 57~78
- 5 Zhan Xuegang, Yao Tianshun. A Chinese Document Classification Method Based on Syntax Analysis. In: Proc. 1998 Intl. Conf. on Chinese Information Processing. China, 1998
- 6 Dong Mingkai, Tian Qijia, Shi Zhongzhi. Web spider based on intelligent agent. In: Proc. 5th World Multiconference on Systemics, Cybernetics and Informatics(SCI 2001) and 7th Intl. Conf. on Information Systems Analysis and Synthesis(ISAS 2001)
- 7 Wang Bo, Tian Qijia, Shi Zhongzhi. Multi-agent environment. In: Proc. 5th World Multiconference on Systemics, Cybernetics and Informatics(SCI 2001) and 7th Intl. Conf. on Information Systems Analysis and Synthesis(ISAS 2001)
- 8 (上接第16页)
- 9 Bolot J C, Turletti T, Wakeman I. Scalable Feedback Control for Multicast Video Distribution in the Internet. In: Proc. of ACM SIGCOMM'94, Sept. 1994. 58~67
- 10 Montgomery T. A loss tolerant rate controller for reliable multicast:[Technical Report NASA-IVV-97-011]. West Virginia University, Aug. 1997
- 11 Sano T, et al. Flow and congestion control for bulk reliable multicast protocols toward coexistence with TCP. RM meeting, Sept. 1997
- 12 Turletti T, Huitema C. Videoconferencing on the Internet. IEEE ACM Transactions on Networking, 1996, 4(3): 340~547
- 13 McCanne S, Jacobson V, Vetterli M. Receiver-Driven Layered Multicast. In: Proc. of ACM SIGCOMM'96, (Stanford, CA), Sept. 1998
- 14 Wu L, Sharma R, Smith B. Thin Stream: An Architecture for Multicasting Layered Video. In: Proc. NOSSDAV'97, (St. Louis, MI), May 1997
- 15 Floyd S, Fall K. Promoting the use of end-to-end congestion control in the Internet. submitted to IEEE/ACM Transaction of networking, Feb. 1998
- 16 Rubenstein D, Kurose J, Towsley D. The Impact of Multicast Layering on Network Fairness. ACM SIGCOMM'99, Cambridge, MA, Sep. 1999
- 17 ViCisano L, Rizzo L, Crowcroft J. TCP-like Congestion Control for Layered Multicast Data Transfer. IEEE INFOCOM'98, (San Francisco, CA), Mar. 1998
- 18 Li X, Paul S, Ammar M. Multi-Session Rate Control for Layered Video Multicast. In: Proc. of Symp. on Multimedia Computing and Networking (MMCN'99), (San Jose, CA), Jan. 1999
- 19 Amir E, McCanne S, Zhang H. An Application Level Video Gateway. ACM Multimedia, 1995(Nov.): 511~522
- 20 Cheung S, Ammar M H, Li X. On the Use of Destination Set Grouping to Improve Fairness in Multicast Video Distribution. In: Proc. of IEEE INFOCOM'96, (San Francisco, CA), Mar. 1996. 553~560
- 21 Xu X, et al. Resilient Multicast for Continuous-Media Applications. In: G. Parulkar, Ed. NOSSDAV'97, Springer, May 1997
- 22 Pejhan S, Schwartz M, Anatssiou D. Error Control Using Retransmission Schemes in Multicast Transport Protocols for Real-Time Media. IEEE/ACM Transactions on Networking, 1996, 4(3): 413~427
- 23 Li X, Paul S, Ammar M H. Layered Video Multicast with Retransmission (LVMR): Evaluation Error Recovery. In: Proc. of NOSSDAV'97, (St. Louis, MI), May 1997
- 24 Albanese A, Luby M. Priority Encoding Transmission. IEEE Trans. For Theory, special issue devoted to doding theory, 1996, 42(6)
- 25 Rubenstein D, Kurose J, Towsley D. Real-time Reliable Multicast Using Proactive Forward Error Correction:[CMPSCI Tech Report 98-19]. University of Massachusetts at Amherst, Mar. 1998
- 26 Nonnenmacher J, Biersack E W, Towsley D. Parity-Based Loss Recovery for Reliable Multicast Transmission. In: Proc. ACM SIGCOMM'97, Cannes, France, Sept. 1997. 289~300